Articulatory Speech Synthesis from Static Context-Aware Articulatory Targets

Anastasiia Tsukanova[1](✉), Benjamin Elie[2], and Yves Laprie[1]

[1] Université de Lorraine, CNRS, Inria, LORIA, 54000 Nancy, France
anastasiia.tsukanova@inria.fr

[2] L2S, CentraleSupelec, CNRS, Université Paris-sud, Université Paris-Saclay,
3 rue Joliot-Curie, 91192 Gif-sur-Yvette, France

Abstract. The aim of this work is to develop an algorithm for controlling the articulators (the jaw, the tongue, the lips, the velum, the larynx and the epiglottis) to produce given speech sounds, syllables and phrases. This control has to take into account coarticulation and be flexible enough to be able to vary strategies for speech production. The data for the algorithm are 97 static MRI images capturing the articulation of French vowels and blocked consonant-vowel syllables. The results of this synthesis are evaluated visually, acoustically and perceptually, and the problems encountered are broken down by their origin: the dataset, its modeling, the algorithm for managing the vocal tract shapes, their translation to the area functions, and the acoustic simulation. We conclude that, among our test examples, the articulatory strategies for vowels and stops are most correct, followed by those of nasals and fricatives. Improving timing strategies with dynamic data is suggested as an avenue for future work.

Keywords: Articulatory synthesis · Coarticulation
Articulatory gestures

1 Introduction

Articulatory speech synthesis is a method of synthesizing speech by managing the vocal tract shape on the level of the speech organs, which is an advantage over the state-of-the-art methods that do not usually incorporate any articulatory information. The vocal tract can be modeled with geometric [2,16,18], biomechanical [1,13] and statistical [9,14] models. The advantage of statistical models is that they use very few parameters, speeding up the computation time. Their disadvantage is that they follow the data a priori without any guidance and do not have access to the knowledge of what is realistic or physically possible. Because of this, to produce correct configurations, they need to be finely tuned.

We were interested in exploring the potential in using quite little, and yet sufficient, static magnetic resonance imaging (MRI) data and implementing one

© Springer Nature Switzerland AG 2018
Q. Fang et al. (Eds.): ISSP 2017, LNAI 10733, pp. 37–47, 2018.
https://doi.org/10.1007/978-3-030-00126-1_4

of the few existing attempts at creating a full-fledged speech synthesizer that would be capable of reproducing the vast diversity of speech sounds.

2 Building an Articulatory Speech Synthesis System

The system is basically made up of three components: the database with the "building blocks" for articulating utterances, the joint control algorithm for the vocal tract and the glottal source, and acoustic simulation. The primary concern of this work are the first two components.

2.1 Dataset

The dataset construction and manipulation were inspired by the work of [3]. We used static MRI data, 97 images collected with a GE Signa 3 T machine with an 8-channel neurovascular coil array. The protocol consisted in a 3D volume acquisition of the vocal tract acquired with a custom modified Enhanced Fast Gradient Echo (EFGRE3D, TR 3.12 ms, TE 1.08 ms, matrix 256 × 256 × 76, with spatial resolution $1.02 \times 1.02 \times 1.0\ mm^3$). Then we selected the mid-sagittal slice in those images. These data captured articulation without phonation: the speaker was instructed to show the position that he would have to attain to produce a particular sound. For vowels, that is the position when the vowel would be at its clearest if the subject were phonating. For consonant-vowel (CV) syllables, that is the blocked configuration of the vocal tract, as if the subject were about to start pronouncing it. The assumption is that such articulation shows the anticipatory coarticulation effects of the vowel V on the consonant C preceding it. There were 13 vowels, 72 CV syllables and 2 semi-vowels in the final dataset. This covers all main phonemes of the French language, but not in all contexts. Each consonant was recorded in the context of the three cardinal vowels and /y/, which is strongly protruded in French. Some intermediate vocalic contexts were added so as to enable the vowel context expansion algorithm to be checked.

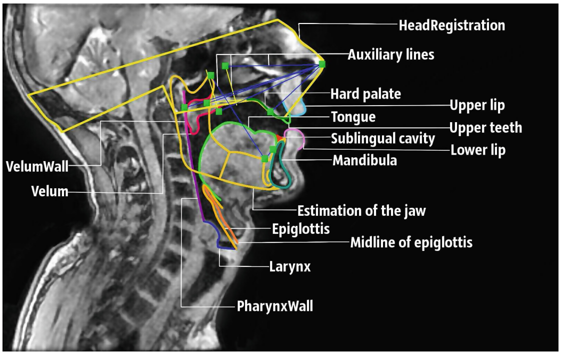

Fig. 1. An example of dataset image annotation *(/a/).*

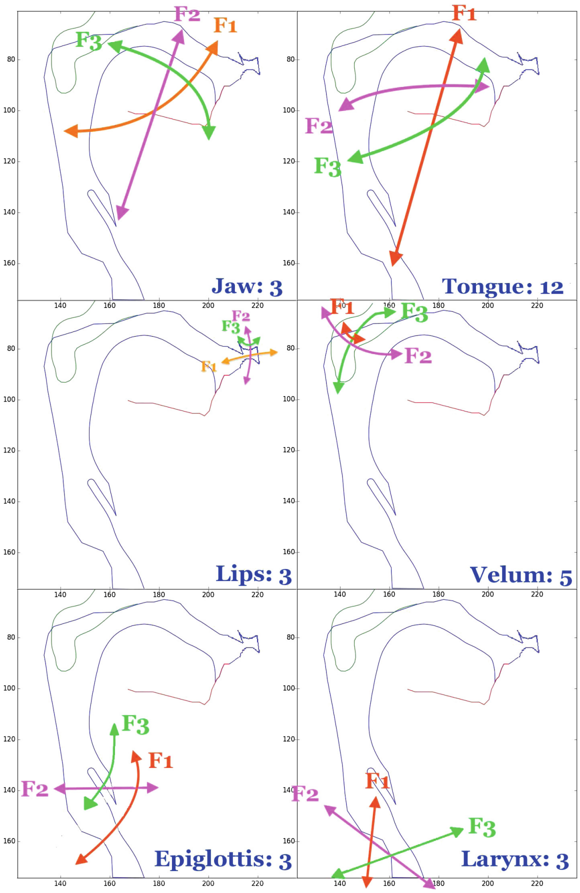

Fig. 2. The PCA-based articulatory model: curve change directions encoded in the first three factors of each articulator (the jaw, the tongue, the lips, the epiglottis, the larynx).

Expanding the Dataset. Since the collected French phonemic dataset was limited, we needed to expand it to cover other contexts as well. We used the notion of the cardinal vowels—/a/, /i/, /u/ and /y/,—assuming that /a/, /i/, /u/ and /y/ represent the most extreme places of vowel articulation, and since then any other vowel articulation can be expanded as a combination of its /a/, /i/ /u/ and /y/ "components". Having captured the C+/a/, C+/i/, C+/u/ and C+/u/context for all consonants C and all non-cardinal vowels V on their own, we were able to estimate the missing C+V samples:

- We projected the vowel V articulatory vector (from $\mathbb{R}^{29}$) onto the convex hull of the /a/, /i/, /u/ and /y/ vectors.
- Assuming that the linear relationship between the C+V vector and the C+/a/, C+/i/, C+/u/ and C+/y/ vectors is the same as the one between V

and /a/, /i/, /u/ and /y/, we estimated C+V from C+/a/, C+/i/, C+/u/ and C+/y/ using the coefficients from the projection of V onto the convex hull of /a/, /i/, /u/ and /y/.

We also estimated the neutral C configuration, the one without any anticipatory effects, as the average of C+/a/, C+/i/, C+/u/ and C+/y/.

Finally, we assumed that the voiced and unvoiced consonants did not have any differences in the articulation.

Articulatory Model. After manually annotating the captures as shown in Fig. 1 we applied a principal-component-analysis (PCA)-based model on the articulator contours [10–12]. We paid special attention to the interaction between articulators and the relevance of deformation modes. Moreover, articulators other than the jaw, tongue and lips are often neglected and modeled with insufficient precision, whereas they can strongly influence acoustics at certain points in the vocal tract. Here are two examples. The position of the epiglottis, which is essentially a cartilage, is likely to modify the geometry of the lower part of the vocal tract by adding an artificial constriction disturbing all the acoustics. It is therefore important to model its deformation modes and interactions with other articulators correctly. In the same way, the velum plays an important role both in controlling the opening of the velopharyngeal port, and in slightly modifying the oral cavity to obtain resonant cavities that give the expected formants of vowels. The acoustic tests we have carried out show in particular that the velum makes it possible to better control the balance between the two cavities necessary for the realization of /u/ and /i/.

Regarding the tongue, PCA was applied on the contours delineated from images. Deformation modes are likely to be impacted by delineation errors. In the case of the tongue, these errors are marginal, or at least give rise to deformation modes coming after the genuine deformations whose amplitude is bigger. On the other hand, the width of epiglottis and/or velum is small on the images, and the errors of delineation, whether manual or automatic, are of the same order of magnitude as genuine deformations. Consequently, PCA applied without precaution will mix both types of deformation. To prevent the apparition of these spurious deformation components the epiglottis was approximated as a thick curve, and only the centerline of epiglottis was analyzed. As a matter of fact, the centerline was determined after delineation of all the epiglottis contours, and the width was set as the average width of all these contours in the upper part where the two epiglottis edges are clearly visible (see Fig. 3). The height of the upper part (where both contours are visible) is adjusted by hand to fit the contours extracted from images. The centerline is approximated as a B-spline and represented by its control points P_l ($0 \leq l < M$ where M is the number of control points) in the form of a two-coordinate vector, and the reconstruction of the epiglottis from the centerline amounts to draw a line at a distance of half the width from the centerline.

The influence of delineation errors is very similar for the velum, which is a fairly fine structure not always well marked on MRI images because it moves

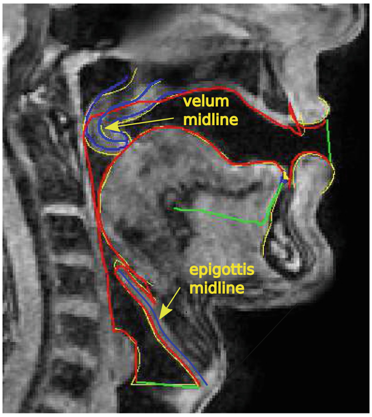

Fig. 3. Epiglottis and velum centerlines reconstructed by the model (solid blue lines). Reconstructed vocal tract is represented by solid red lines. The vocal tract input and output are represented in solid green lines. All these contours are superimposed onto the contours (represented as solid yellow lines) delineated from the image. (Color figure online)

quickly. As for epiglottis we used the centerline and a fairly simple reconstruction algorithm. However, PCA was not applied directly on the control point of the splines because the velum can roll up on itself. This particularity does not lend itself well to the direct use of PCA, which results in the emergence of linear components not appropriate in this case. The centerline is therefore broken down into a series of segments of the same length. Each segment articulates with its predecessor and the first point is fixed. The centerline is then defined as the vector of angles between two consecutive segments. In this way PCA can be applied effectively to velum and gives rise to relevant deformation modes.

The architecture and general organization of the articulatory model are based on the dependency links between the articulators. The main articulator is the jaw which is represented by 3 parameters to get a complete and accurate control. Its geometrical contribution is subtracted from the tongue contours before the application of PCA because tongue is directly attached to the mandible. The tongue is represented by 12 parameters in order to obtain a sufficient precision for the realisation of consonant constrictions. The lips are represented by 3 parameters. Unlike the tongue, the interactions between lips and jaw are more complex. For this reason we subtract the correlation between jaw and lips before applying PCA. The larynx is considered to be independent of the jaw and is represented by 3 parameters to control ist orientation and vertical position. In the same way the velum is considered as an articulator independent of the others. It is analyzed as explained above and is represented by 3 parameters. The epiglottis is the articulator that is subject to the greatest number of influences: the jaw via the tongue, the tongue itself and the larynx. These influences are subtracted by applying a multiple regression to the epiglottis centreline before applying CPA. Analysis of the variance shows that the various influences on the epiglottis account for most of its deformations. Its intrinsic deformation are represented by 3 parameters.

In total these parameters form a vector from $\mathbb{R}^{27}$ (see Fig. 2 for major parameter contributions to the articulator shape). Since the model uses PCA, the zero configuration should correspond to the central position as identified in the dataset, and small changes in the parameter space within a certain neighborhood of zero should correspond to small changes (in terms of distance and shape, not in terms of the resulting acoustics) in the curves. A clipping algorithm is used to solve problems of collision between articulators, i.e. essentially between the tongue and palate. So the model's behavior is not entirely linear.

2.2 Strategies for Transitioning Between the Articulatory Targets

The dataset provided static images capturing idealistic, possibly over-articulated, targets for consonants anticipating particular vowels, whereas the goal was to be also able to deal with consonant clusters and consonants that would not anticipate any vowel at all—for example, due to their ultimate position in a rhythmic phrase. So, in our context, to establish a transitioning strategy would mean three things:

- Choose the building blocks: identify the articulatory target for each phoneme in a phrase. It can either be what was captured in the dataset (a vowel or a consonant assuming vocalic anticipation) or an estimation of what the dataset was missing (missing phonemes, such as voiced consonants, missing contexts or the absence of any context). A consonant cannot anticipate multiple phonemes, nor can vowels anticipate anything due to the restrictions of the dataset at our disposal.
- Decide when — and whether — the articulatory target should be attained.
- Decide how to generate the articulatory positions between the target ones.

Our basic assumption was that by default, consonants anticipate the next coming vowel. However, it would be unrealistic to assume it happens in all cases. This is why we imposed several restrictions on the anticipatory effect:

- Temporal: no coarticulatory effect if the anticipated phoneme is more than 200 ms ahead;
- Spatial: if there is any movement scheduled between the anticipated vowel, the phoneme in question negates the effect. For example, consider such sequence as /lki/: after /l/, the tongue needs to move backward to produce /k/ before coming back forward for /i/. In this situation, our algorithm does not allow the /l/ to anticipate the coming /i/;
- Categorical: it is not possible to anticipate a vowel more than 5 phonemes ahead, and this restriction becomes stricter if it applies across syllable boundaries.

For vowels, there is also a model of target undershoot.

Having established the articulatory targets, the question is how to transition between them. We have tested out three strategies for interpolation between the target vectors:

- Linear: the interpolation between the target vectors is linear;
- Cosine;
- Complex: transitions are done with cosine interpolation, but the timing varies by the articulators. The critical ones reach for their target position faster than the others, while those articulators whose contribution to the resulting sound intelligibility is not as large move slower (for example, the tongue can be in a number of positions for the sound /b/, but the lips have to come into contact). Besides, the articulators composed of heavier tissues (such as the tongue back) move slower than the light and highly mobile ones (such as the lips).

2.3 Obtaining the Sound

Each vocal tract position was encoded in an area function. They were obtained by the algorithm of [7] with coefficients adapted by S. Maeda and Y. Laprie. These parameters only depend on the position in the vocal tract between the glottis and the lips. The transition from the sagittal view to the area function has given rise to several works which contradict each other slightly ([17] and [15]) and it is therefore clear that the determination of the area function will have to take into account the dynamic position of the articulators in the future.

We used an acoustic simulation system implemented by [4] to obtain sound from the area functions and supplementary control files: glottal opening and pitch control.

Glottal opening was modeled by using external lighting and sensing photoglottography (ePGG) measurements [8]. Within the model, glottis is at its most closed position when producing vowels, nasals and the liquid sound /l/, and momentarily reaches its most open one when producing voiceless fricatives and stops. Voiced fricatives and stops also create peaks in glottal opening, but not as high.

There was no need to model voicing (high-frequency oscillations of low amplitude superimposed onto the glottal opening waves) since the vocal folds operated by the glottal chink model [4,5] are self-oscillating.

3 Evaluation

Each step in the system was evaluated on its own, and afterwards the synthesis results were evaluated visually, acoustically and perceptually. Since the objective of the work was rather to have a fully functional algorithm that produces reasonably realistic movements and sounds rather than to obtain high-quality speech, a more rigorous evaluation, such as a quantitative comparison to the dynamic data on articulatory trajectories, is still an avenue of future work.

3.1 The Articulatory Model and the Trajectories

One peculiarity of the dataset and therefore of the model was the fact that it used only the sagittal section of the speaker's vocal tract. While full three-dimensional models can capture the full geometry of the vocal tract with such phenomena as lateral phonemes (e.g. /l/), two-dimensional models get the benefit of faster computation time and overall simplicity, but irreversibly lose the spatial information.

In general, the articulatory model captured vocal tract positions correctly or with no critical errors, and some adjustments could be necessary only at the points of constriction, since on its own the model did not impose much control over them. This is a disadvantage brought by the nature of the articulatory model: choosing to operate at the level of articulators rather than the resulting vocal tract geometry.

As for the movements, for now we can say that they were reasonable and the coarticulation-affected targets guided the articulators to the positions necessary to produce a particular utterance. One key point here is the timing strategy. Rule-based timing strategy seems to be very rigid for the dynamic nature of speech; it would be more natural to follow speech production processes in humans and to guide the synthesis with the elicited sound or the speaker's expectation—based on their experience—on what this sound will be. We plan to evaluate the transitions with new dynamic MRI data.

3.2 Glottal Opening Control

The algorithm for the glottis opening successfully allowed to distinguish between vowels and consonants. Distinguishing between voiced and voiceless consonants, though, stays a point for improvement, as well as well-coordinated control over the glottis and the vocal tract.

3.3 The Synthesized Sound

Vowels and stops were the most identifiable and correct, although sometimes some minor adjustments in the original data were necessary to obtain formants close to the reference values. When compared to human speech, the formant transitions within the suggested strategies sometimes occurred too fast and sometimes too slowly; again, this highlights the utmost importance of realistic timing strategies. Figure 4 shows an example of the synthesis when it is guided by real timing: /aʃa/ as produced by the system and as uttered by a human. The high-frequency contributions in /ʃ/, not appearing in the human sample, are due to the acoustic simulation. The noise of /ʃ/ is at the correct frequencies, but with a bit different energy distribution, probably because of differences in articulation or in the area functions. There is also an acoustic artifact between /ʃ/ and /a/, which means that more work is necessary on liaising the vocal tract and the source control.

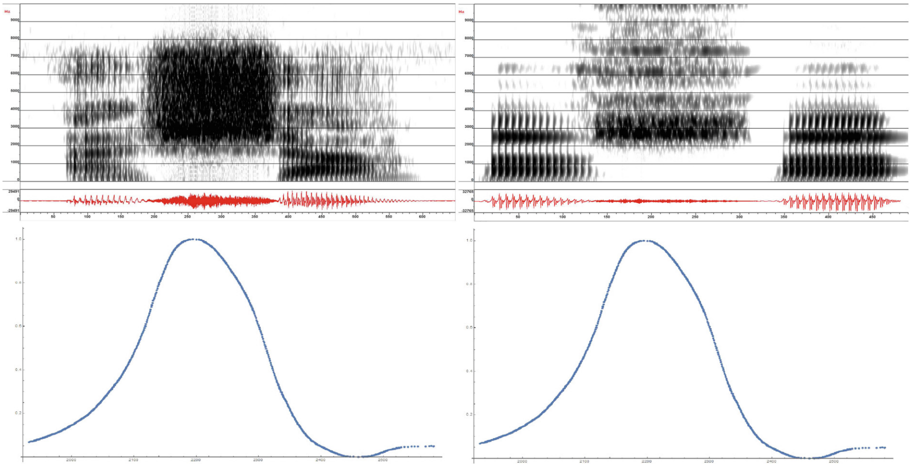

Fig. 4. An example of a human's utterance of /aʃa/ (left) and its synthesis (right) along with the glottal closure control as copied from the EPGG data (below). Phoneme durations are aligned.

4 Conclusion

Regarding speech as a process of transitioning between context-aware targets is an interesting approach that can be connected with the mental processes of speech production: to allow the others to perceive the necessary acoustic cue, the speaker needs to come close enough to the associated articulatory goal. The important difference between a real speaker and the algorithm is the fact that the algorithm solves a static problem, laid out in full; it needs to hit particular targets in a given order. As for humans, we solve a dynamic problem, and coarticulation is not something we put in its definition; rather, coarticulation is our means to make the problem of reaching too many targets in a too short period of time solvable.

The statistically derived articulatory model encodes complicated shapes of the articulators in only 29 parameters, sometimes struggling at the constrictions because of the inherent—and intentional—lack of control over the resulting geometry of the vocal tract.

Those shapes of the articulators change in time according to the produced trajectories of the vocal tract, and those are phonetically sound. Whether there are any important differences between the produced transitions and the ones in real speech, needs to be verified with actual dynamic data.

After the aspect of *how* the articulators move we need to consider *when*. The timing strategies, currently rule-based, apparently need to be extracted from dynamic data, and we can use the approaches by [6] for that.

A closer, intertwined interaction with the acoustic simulation unit—such as guidance on how to navigate between the area functions at the level of separate acoustic tubes and improved control over the glottal opening—could improve the results for consonants.

Acknowledgments. The data collection for this work benefited from the support of the project ArtSpeech of ANR (Agence Nationale de la Recherche), France.

References

1. Anderson, P., Harandi, N.M., Moisik, S., Stavness, I., Fels, S.: A comprehensive 3D biomechanically-driven vocal tract model including inverse dynamics for speech research. In: Sixteenth Annual Conference of the International Speech Communication Association (2015)
2. Birkholz, P., Jackèl, D., Kröger, B.J.: Construction and control of a three-dimensional vocal tract model. In: Proceedings of International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2006), pp. 873–876 (2006)
3. Birkholz, P.: Modeling consonant-vowel coarticulation for articulatory speech synthesis. PloS one **8**(4), e60603 (2013)
4. Elie, B., Laprie, Y.: Extension of the single-matrix formulation of the vocal tract: consideration of bilateral channels and connection of self-oscillating models of the vocal folds with a glottal chink. Speech Commun. **82**, 85–96 (2016)
5. Elie, B., Laprie, Y.: A glottal chink model for the synthesis of voiced fricatives. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5240–5244. IEEE (2016)
6. Elie, B., Laprie, Y., Vuissoz, P.A., Odille, F.: High spatiotemporal cineMRI films using compressed sensing for acquiring articulatory data. In: Eusipco, Budapest, pp. 1353–1357, August 2016
7. Heinz, J.M., Stevens, K.N.: On the relations between lateral cineradiographs, area functions and acoustic spectra of speech. In: Proceedings of the 5th International Congress on Acoustics, p. A44 (1965)
8. Honda, K., Maeda, S.: Glottal-opening and airflow pattern during production of voiceless fricatives: a new non-invasive instrumentation. J. Acoust. Soc. Am. **123**(5), 3738–3738 (2008)
9. Howard, I.S., Messum, P.: Modeling the development of pronunciation in infant speech acquisition. Motor Control **15**(1), 85–117 (2011)
10. Laprie, Y., Busset, J.: Construction and evaluation of an articulatory model of the vocal tract. In: 19th European Signal Processing Conference - EUSIPCO-2011. Barcelona, Spain, August 2011
11. Laprie, Y., Vaxelaire, B., Cadot, M.: Geometric articulatory model adapted to the production of consonants. In: 10th International Seminar on Speech Production (ISSP). Köln, Allemagne, May 2014. http://hal.inria.fr/hal-01002125
12. Laprie, Y., Elie, B., Tsukanova, A.: 2D articulatory velum modeling applied to copy synthesis of sentences containing nasal phonemes. In: International Congress of Phonetic Sciences (2015)
13. Lloyd, J.E., Stavness, I., Fels, S.: ArtiSynth: a fast interactive biomechanical modeling toolkit combining multibody and finite element simulation. In: Payan Y. (eds.) Soft Tissue Biomechanical Modeling for Computer Assisted Surgery, pp. 355–394. Springer, Berlin (2012).https://doi.org/10.1007/8415_2012_126

14. Maeda, S.: Compensatory articulation during speech: Evidence from the analysis and synthesis of vocal-tract shapes using an articulatory model. In: Hardcastle, W., Marchal, A. (eds.) Speech Production and Speech Modelling, pp. 131–149. Kluwer Academic Publisher, Amsterdam (1990)
15. McGowan, R., Jackson, M., Berger, M.: Analyses of vocal tract cross-distance to area mapping: an investigation of a set of vowel images. J. Acoust. Soc. Am. **131**(1), 424–434 (2012)
16. Öhman, S.: Coarticulation in VCV utterances: spectrographic measurements. J. Acoust. Soc. Am. **39**(1), 151–168 (1966)
17. Soquet, A., Lecuit, V., Metens, T., Demolin, D.: Mid-sagittal cut to area function tranformations: direct measurements of mid-sagittal distance and area with MRI. Speech Commun. **36**(3–4), 169–180 (2002)
18. Story, B.: Phrase-level speech simulation with an airway modulation model of speech production. Comput. Speech Lang. **27**(4), 989–1010 (2013)